# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-266393

(43)Date of publication of application : 22.09.1994

(51)Int.Cl. G10L 5/06

G10L 3/00

G10L 3/00

(21)Application number : 05-051897

(22)Date of filing : 12.03.1993

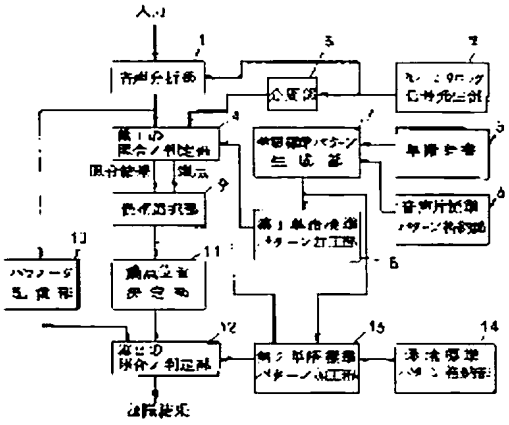
(71)Applicant : MATSUSHITA ELECTRIC IND CO LTD

(72)Inventor : KIMURA TATSUYA
KUWANO HIROYASU
WATANABE TAISUKE
HIRAOKA SEIJI

(54) SPEECH RECOGNITION DEVICE

(57)Abstract:

PURPOSE: To reduce the calculation quantity while securing the recognition performance by narrowing down word candidates previously by collation arithmetic, and then collating the narraowed-down candidates by using data which are not thinned out.



CONSTITUTION: A 2nd collation/decision part 12 collates parameters and word standard patterns stored in a parameter storage part 10 as to word candidates and a collation section without a thinning-out process while an end point is fixed and then outputs the word candidate which gives the best collation result as a recognition result. The collation section supplied to the 2nd collation/ decision part 12 is made longer than an actual speech section, so the word standard parameters used for the collation are used after performing a processing for connecting an environment standard pattern stored in an environment standard pattern storage part 14 to both ends of the word standard pattern obtained by a word standard pattern generation part 7 is performed by a 2nd word standard pattern processing part 13.

---

LEGAL STATUS

| | |
|---|---|
| [Date of request for examination] | 12.10.1999 |
| [Date of sending the examiner's decision of rejection] | |
| [Kind of final disposal of application other than the examiner's decision of rejection or application converted registration] | |
| [Date of final disposal for application] | |
| [Patent number] | 3428058 |
| [Date of registration] | 16.05.2003 |
| [Number of appeal against examiner's decision of rejection] | |

[Date of requesting appeal against
examiner's decision of rejection]
[Date of extinction of right]

---

CLAIMS

---

[Claim(s)]
[Claim 1] A voice-analysis means which is the base unit of analysis of an input sound signal to analyze for every frame and to extract an analysis parameter, A frame clock generation means to utter the timing signal of an analysis frame, A dividing means to carry out dividing of the above-mentioned frame clock by the predetermined division ratio, and to acquire a dividing clock signal, A voice piece standard-pattern storing means to store the voice piece standard pattern constituted by the sequence of the partial standard pattern expressing some of word dictionaries which wrote the word by the sequence of the notation showing a voice piece, and piece data of voice created using the voice data which many men uttered beforehand, A word standard-pattern generation means to obtain the standard pattern of a word by connecting the above-mentioned voice piece standard pattern according to the content of a notation of the above-mentioned word dictionary, The 1st word standard-pattern processing means which creates a data infanticide word standard pattern by thinning out a part of feature-parameter sequence which constitutes the above-mentioned word standard pattern, The partial distance which is the distance between the feature parameters and the partial standard patterns of the above-mentioned data infanticide word standard pattern which are obtained from the above-mentioned analysis parameter in a simultaneous point whenever it receives the above-mentioned dividing clock signal is computed. By accumulating the partial distance between the above-mentioned data infanticide word standard patterns already called for from the event concerned and the feature-parameter sequence before it The start edge location which accompanies the minimum distance and it to the input of the data infanticide word standard pattern at the time of assuming the event concerned to be the termination of a word is obtained. 1st collating/judgment means which combines the minimum above-mentioned distance with the above-mentioned start edge location, and updates it for every word for every above-mentioned dividing clock, A candidate word selection means to be at the termination event of input voice and to obtain a predetermined number candidate word in order with a small distance value by

comparing the distance over the standard pattern of all the words for recognition mutually, The endpoint positioning means which determines the section which certainly includes the voice section from the start edge which accompanies the candidate word chosen by the above-mentioned candidate selection means, and a termination candidate group, A parameter storage means to memorize the above-mentioned analysis parameter over all the input sections, An environmental pattern storing means to store the environmental standard pattern beforehand created from the acoustic signal of the sections other than voice, The 2nd word standard-pattern processing means which connects the above-mentioned environmental pattern before and after the above-mentioned word standard pattern, and creates a word standard pattern with an environmental pattern, It computes, when partial distance accumulates the distance between the parameter sequences in the section determined by the above-mentioned endpoint positioning means stored in the word standard pattern with an environmental pattern and the above-mentioned parameter storage means corresponding to the word candidate group chosen by the above-mentioned word candidate selection means. The voice recognition unit which consists of the 2nd collating/judgment means which outputs the word candidate who acquired the distance value with the smallest value by carrying out the mutual comparison of the distance acquired for every above-mentioned candidate word as a recognition result.

[Claim 2] The voice recognition unit according to claim 1 characterized by simplifying the count in calculation and word collating of partial distance in processing of 1st collating/judgment means using inter-frame length.

[Claim 3] The 1st word standard-pattern merge means which creates a data infanticide merge word standard pattern by packing into one the partial standard pattern of the data infanticide word standard pattern created with the 1st word standard-pattern processing means by making the same multiple frame into a group is added. The partial distance count section which computes the partial distance which is the distance between the feature parameters and the partial standard patterns of the above-mentioned data infanticide merge word standard pattern which are obtained from the analysis parameter in a simultaneous point whenever 1st collating/judgment means receives a dividing clock signal, The representation partial distance selection section which compares said partial distance with the partial distance at the front [ event / concerned ] event, and makes the one where distance is smaller representation partial distance, The distance accumulation section which accumulates the representation partial distance between the above-mentioned data

infanticide word standard patterns already called for from the event concerned and the feature-parameter sequence before it, The start edge location which accompanies the minimum distance and it to the input of the data infanticide word standard pattern at the time of assuming the event concerned to be the termination of a word is obtained. The voice recognition unit according to claim 1 characterized by having the judgment section which combines the minimum above-mentioned distance with the above-mentioned start edge location, and updates it for every word for every above-mentioned dividing clock, and simplifying the count in calculation and word collating of partial distance.

[Claim 4] Partial distance is a voice recognition unit according to claim 1 to 3 which computes using a statistical interval scale and is characterized by the above-mentioned statistical interval scale being an interval scale based on a-posteriori probability.

[Claim 5] Partial distance is a voice recognition unit according to claim 1 to 3 which computes using a statistical interval scale and is characterized by the above-mentioned statistical interval scale being a primary discriminant based on a-posteriori probability.

---

---

DETAILED DESCRIPTION

---

[Detailed Description of the Invention]

[0001]

[Industrial Application] This invention relates to the approach of speech recognition of making a machine recognizing human being's voice.

[0002]

[Description of the Prior Art] Although there are an approach for a specified speaker and an approach for an unspecified speaker in speech recognition, unspecified speaker recognition is targetted especially for this invention. As an example of the approach for an unspecified speaker, the example based on Japanese Patent Application No. No. 314248 [ three to ] is explained, referring to drawing 9 .

[0003] In drawing 7 61 the feature-parameter extract section and 63 for the sonagraphy section and 62 A voice section detecting element, The word dictionary in which two or more frame buffers and 65 described all the words for recognition in the voice piece standard-pattern storing section, and 64 described 66 along the voice

piece, The word standard-pattern generation section which generates the word standard pattern of the vocabulary for recognition when 67 chooses and connects a voice piece standard pattern according to the list of a voice piece, The partial distance count section which finds the partial distance of the input vector and the partial pattern of the voice for recognition with which 68 was formed with two or more frames by the statistical interval scale based on a-posteriori probability, It is the judgment section which makes a recognition result the voice name which the distance accumulation section which finds the distance of input voice and a word standard pattern by accumulating partial distance over the whole voice, and 610 make accumulation distance as the path judging section while 69 shifts an input frame, and 611 makes min.

[0004] The sonagraphy section 61 carries out the AD translation of the input signal, and is fixed time amount length (it is called a frame.). In this conventional example, it analyzes to every 10ms. In the feature-parameter extract section 62, a feature parameter is extracted based on the output of the sonagraphy section 61. The voice section detecting element 63 detects the start edge of input signal voice, and termination. Although the method of detecting the voice section is easy and is common, what kind of approach may be used. [ of the method of using audio power ] Moreover, about this, it mentions later using the approach of word spotting which performs a collating operation, assuming the entire interval of an input to be an endpoint, without performing voice section detection. Two or more frame buffers 64 are parts which form the input vector which unifies the feature parameter of the frame of the neighborhood of each frame, and is used for pattern matching (partial matching). The standard pattern of a voice piece is stored in the voice piece standard-pattern storing section 65 as association of a partial pattern. The link information of a voice piece is described by the word dictionary 66 for every word to recognize. The voice piece connection section 67 reads and connects the voice piece standard pattern stored in the voice piece standard-pattern storing section 65 according to this voice piece link information. In the partial distance count section 68, the distance between a word standard pattern and two or more frame buffers (partial distance) is calculated. The distance accumulation section 69 accumulates the partial distance over each word, and asks for the similarity to the whole word. The path judging section 610 chooses the path from which accumulation distance becomes min. The judgment section 611 is outputted in quest of the word which gives the minimum value of accumulation distance.

[0005] Next, the case where the word-spotting method do not perform voice section

detection is used is explained. Since the voice section detection generally weakened at a noise does not need to be used for the advantage of the word-spotting method, it is that a recognition system strong against a noise is realizable. Since voice section detection is not performed in the case of the word-spotting method, a collating operation is performed about sufficient long section containing voice. That is, using a collating initiation event as the audio start edge, and carrying out a collating operation for a collating termination event as audio termination like [ in the case of performing voice section detection ], does not have semantics. By the word-spotting method, the collating score to a word standard pattern is computed by assuming an audio always edge about all the input sections.

[0006]

[Problem(s) to be Solved by the Invention] The speech recognition for unspecified speakers with a high precision is possible for the approach explained in the conventional example using positively the information on "a neighboring inter-frame time motion", and by using a statistical interval scale. Moreover, since it is the approach of connecting a voice piece, implementation of the high recognition equipment of the versatility in which lexical modification is possible only by rewriting a word dictionary is possible. Furthermore, since precise voice section detection becomes unnecessary by performing word spotting, recognition equipment strong against a noise is realizable.

[0007] However, since this approach was finding partial distance about all the input sections and the entire interval of the standard pattern of a word in addition to the number of dimension of a feature parameter being large since the analysis parameter of the section (multiple frame) with width of face including the neighboring frame of the frame made into the feature parameter is used, although it used the linearity discernment type for count of partial distance, it had the trouble that there was still much computational complexity. Moreover, when word spotting was used, there was a problem of "partial matching" that a certain word matched with a part of other words, and served as incorrect recognition like the example of "Fujiidera" and "Fuji."

[0008]

[Means for Solving the Problem] In order to solve the problem described above in this invention A voice-analysis means which is the base unit of analysis of an input sound signal to analyze for every frame and to extract an analysis parameter, A frame clock generation means to utter the timing signal of an analysis frame, A dividing means to carry out dividing of the above-mentioned frame clock by the predetermined division ratio, and to acquire a dividing clock signal, A voice piece standard-pattern storing

means to store the voice piece standard pattern constituted by the sequence of the partial standard pattern expressing some of word dictionaries which wrote the word by the sequence of the notation showing a voice piece, and voice piece data created using the voice data which many men uttered beforehand, A word standard-pattern generation means to obtain the standard pattern of a word by connecting the above-mentioned voice piece standard pattern according to the content of a notation of the above-mentioned word dictionary, The 1st word standard-pattern processing means which creates a data infanticide word standard pattern by thinning out a part of feature-parameter sequence which constitutes the above-mentioned word standard pattern, The partial distance which is the distance between the feature parameters and the partial standard patterns of the above-mentioned data infanticide word standard pattern which are obtained from the above-mentioned analysis parameter in a simultaneous point whenever it receives the above-mentioned dividing clock signal is computed. By accumulating the partial distance between the above-mentioned data infanticide word standard patterns already called for from the event concerned and the feature-parameter sequence before it The start edge location which accompanies the minimum distance and it to the input of the data infanticide word standard pattern at the time of assuming the event concerned to be the termination of a word is obtained. 1st collating/judgment means which combines the minimum above-mentioned distance with the above-mentioned start edge location, and updates it for every word for every above-mentioned dividing clock, A candidate word selection means to be at the termination event of input voice and to obtain a predetermined number candidate word in order with a small distance value by comparing the distance over the standard pattern of all the words for recognition mutually, The endpoint positioning means which determines the section which certainly includes the voice section from the start edge which accompanies the candidate word chosen by the above-mentioned candidate selection means, and a termination candidate group, A parameter storage means to memorize the above-mentioned analysis parameter over all the input sections, An environmental pattern storing means to store the environmental standard pattern beforehand created from the acoustic signal of the sections other than voice, The 2nd word standard-pattern processing means which connects the above-mentioned environmental pattern before and after the above-mentioned word standard pattern, and creates a word standard pattern with an environmental pattern, It computes, when partial distance accumulates the distance between the parameter sequences in the section determined by the above-mentioned endpoint positioning means stored in

the word standard pattern with an environmental pattern and the above-mentioned parameter storage means corresponding to the word candidate group chosen by the above-mentioned word candidate selection means. 2nd collating/judgment means which outputs the word candidate who acquired the distance value with the smallest value as a recognition result is established by carrying out the mutual comparison of the distance acquired for every above-mentioned candidate word.

[0009]

[Function] After this invention narrows down a word candidate beforehand by the collating operation using the data thinned out by each above-mentioned means division dividing means and the 1st word standard-pattern processing means The 1st operation effectiveness of aiming at the cutback of computational complexity while securing the recognition engine performance by performing collating using the data which do not cull out to the narrowed-down candidate, While realizing recognition strong against a noise by collating including the section of the outside of the voice section by using the above-mentioned word standard pattern with an environmental pattern In the case of word spotting, when a certain word used as a problem collates with the one section of other words, it has the 2nd operation effectiveness of solving the problem of partial matching which incorrect recognition produces.

[0010]

[Example] Hereafter, the 1st example of this invention is explained using a drawing. Drawing 1 shows the configuration of the 1st example of this invention. In drawing 1 the sonagraphy section and 2 1 The frame clock signal generating section, In 3, the dividing section and 4 a word dictionary and 6 for 1st collating/judgment section and 5 The voice piece standard-pattern storing section, 7 -- the word standard-pattern generation section and 8 -- for the parameter storage section and 11, as for 2nd collating/judgment section and 13, the endpoint fixing section and 12 are [ the 1st word standard-pattern processing section and 9 / the candidate selection section and 10 / the 2nd word standard-pattern processing section and 14 ] the environmental standard-pattern storing sections. Next, the actuation is explained.

[0011] The sonagraphy section 1 carries out the AD translation of the input signal, and is fixed time amount length (it is called a frame.). In this example, it analyzes to every 10ms. Linear predictive coding (LPC analysis) is used in the example. The timing of a frame is given by the clock signal which the frame clock signal generating section 2 generates, and this clock signal is supplied to the sonagraphy section 1 and the dividing section 3. The dividing section 3 carries out dividing of the frame clock signal by the predetermined division ratio (this example 2), and outputs a dividing clock signal.

This dividing clock signal is supplied to 1st collating/judgment section 4, and is used for inter-frame length. 1st collating/judgment section 4 performs collating between the analysis parameter which the sonagraphy section 1 outputs, and the word standard pattern generated by the below-mentioned processing by word spotting. About the detail of processing of this part, it mentions later.

[0012] Next, the generation method of the word standard pattern used in 1st collating/judgment section 4 is explained. A recognition vocabulary is expressed along voice piece notations, such as valve flow coefficient and VC, and is stored in the word dictionary 5. A word standard pattern is generated by the word standard-pattern generation section 7 by connecting the voice piece standard pattern stored in the voice piece standard-pattern storing section 6 according to the list of the voice piece notation obtained with reference to the word dictionary 5.

[0013] In addition, about the creation approach of a voice piece standard pattern, it mentions later. Collating in 1st collating/judgment section 4 is performed to the data with which inter-frame length was made for a computational complexity cutback. Therefore, the word standard pattern used in 1st collating/judgment section 4 also needs to give inter-frame length. The 1st word standard-pattern processing section 8 performs processing of this inter-frame length to the word standard pattern obtained in the word standard-pattern generation section 7. Only the predetermined number chooses a word candidate as order with the collating result of all the words from which the candidate selection section 9 was obtained in 1st collating/judgment section 4 to a sufficient collating result. The parameter storage section 10 memorizes the analysis parameter obtained in the voice-analysis section 1 about all the input sections.

[0014] In the endpoint fixing section 11, the information on the start edge obtained along with the above-mentioned word candidate's each and termination is unified, and the collating section for collating in 2nd collating/judgment section 12 is determined. It is determined that this collating section will surely include the voice section. Therefore, the section longer than the actual voice section is obtained. For example, the head of the start edge group obtained along with a word candidate or a location further front is determined as the start edge of the collating section. The same is said of the case of termination and the tail of the termination group obtained along with a word candidate or a further next location is determined as termination of the collating section.

[0015] In 2nd collating/judgment section 12, after collating endpoint immobilization which does not cull out about the above-mentioned word candidate and the collating

section to the word standard pattern obtained according to the parameter memorized by the parameter storage section 10 and the below-mentioned processing, the word candidate who gives the best collating result is outputted as a recognition result. Since the collating section longer than the actual voice section given to 2nd collating/judgment section 12 is taken as above-mentioned, what performed processing which connects the environmental standard pattern stored in the ends of the word standard pattern obtained by the word standard-pattern generation section 7 by the 2nd word standard-pattern processing section 13 at the environmental standard-pattern storing section 14 is used for the word standard pattern used for collating. This environmental standard pattern is created from the pattern of the noise signal with which recognition equipment is used beforehand, for example.

[0016] Next, the content of processing performed in 1st collating/judgment section 4 and 2nd collating/judgment section 12 is explained in detail. It is the point which is collating endpoint immobilization which the latter gives the collating section beforehand to performing the collating operation about the data which do not cull out in the latter, and the former performing word spotting by collating of the endpoint free-lancer who does not give the section of collating to both difference carrying out the collating operation about the data with which the former carried out inter-frame length. Since fundamental views, such as a feature parameter used for other collatings and an interval scale to be used, are the same, the detail of collating processing explains the part which describes 1st collating/judgment section 4, and has a difference each time.

[0017] The detailed block diagram in which drawing 2 shows the flow of processing of 1st collating/judgment section 4, and drawing 3 are the block diagrams of 2nd collating/judgment section 12. Although explanation is mainly given using drawing 2 , drawing 3 and drawing 3 are referred to if needed. Moreover, what has a the same name has the same function about each component of drawing 2 and drawing 3 . For two or more frame buffers and 22, as for the distance accumulation section and 24, in drawing 2 and drawing 3 , the partial distance count section and 23 are [ 21 / the path judging section and 25 ] the judgment sections.

[0018] In drawing 2 , two or more frame buffers 21 are parts which form the input vector which unifies the feature parameter of the frame of the neighborhood of the i-th frame, and is used for pattern matching (partial matching). The input vector in the i-th frame [0019]

[External Character 1]

$X_i$

[0020] ** -- it is expressed as follows.

[0021]

[Equation 1]

$$X_i = (x_{i-L1}, x_{i-L1+m}, \cdots, x_i, \cdots, x_{i+L2})$$

[0022] This is the vector which unified the feature parameter of i-L1 - two i+L every m frames. L1=L2=3 and m= 2 If it carries out, the number of dimension of Xi will be set to x(p+2) {(L1+L2+1) /m+1} =12x4=48. When m takes two or more values, it is equivalent to thinning out a frame and forming an input vector. The voice piece standard-pattern storing section 6 is a part which has stored the standard pattern of a voice piece as association of a partial pattern. The voice piece standard-pattern creating method is explained a little to a detail here.

[0023] the audio element used as a base unit of speech recognition with the [voice piece standard-pattern creation approach] voice piece -- it is -- as a class -- a phoneme, syllable (valve flow coefficient), and semitone -- knot (VC, valve flow coefficient) and vowel - a consonant -- there is a - vowel chain (VCV) etc. In addition, C means a consonant and V means a vowel. The following explanation explains the case where syllable (valve flow coefficient) is used as a class of voice piece as an example.

[0024] For example, the standard pattern of voice piece / sa/is created with the following means.

(1) Start /sa/and the uttered part from the voice data which many men uttered (suppose that the 100-piece sample is started).

(2) Investigate 100 persistence time distribution of /sa/and ask for 100 mean-time length JS.

(3) Discover the sample of the time amount length of JS out of 100 pieces. When there are two or more samples, the average of two or more samples is calculated for every frame. Thus, the called-for representation sample [0025]

[External Character 2]



[0026] [0027]

[Equation 2]

$$S_j = (s_{j-L1}, s_{j-L1+m}, \cdots, s_j, \cdots, s_{j+L2}) \ (j=1, \cdots, JS)$$

[0028] It carries out. here -- sj -- the parameter vector per frame -- it is -- an analysis parameter -- the same -- the LPC cepstrum multiplier of 11 pieces, and difference -- it consists of power.
(4) Perform pattern matching between each sample for 100 pieces (several 1), and a representation sample (several 2), and ask for the inter-frame response relation between the frame of a representation sample, and each sample for 100 pieces (a most similar frame comrade is matched). In addition, if the technique of a dynamic programming is used, it can ask for inter-frame response relation efficiently.
(5) Start the partial vector of the form of (several 1) from each sample for 100 pieces corresponding to each frame (j=1-JS) of a representation sample. Since it is easy l1=l2=3 and m= 1 It carries out.
[0029] It is the partial vector of the n-th sample among the data for 100 pieces equivalent to the j-th frame of a representation sample [0030]
[Equation 3]

$$X^n_j = (x^n_{j-L1},\ x^n_{j-L1+m},\ \cdots,\ x^n_j,\ \cdots,\ x^n_{j+L2})$$

[0031] It carries out. Here shows that j is a frame corresponding to the j-th frame of the inside of the n-th sample of same word / sa/, and a representation vector. In this example, it is a 48-dimensional vector (n=1-100).
(6) 100 [0032]
[External Character 3]

$X^n_j$

[0033] ******** [0034]
[External Character 4]

$\mu_j$

[0035] (48 dimensions) and a covariance matrix [0036]
[External Character 5]

$W_j$

[0037] (48x48 dimensions) are searched for (j=1-JS). As for the average and a covariance matrix, only the number JS of standard frame length will exist (however, it is not necessary to necessarily create these to all frames). You may thin out and create. Above-mentioned (1) - (6) It is [0038] also to voice pieces other than a voice piece / sa by the same procedure.

[External Character 6]
$\mu_j$ 、 $W_j$

[0039] ********. It is the moving average [0040] to all the sample data to all the voice sections.
[External Character 7]
$\mu_x$

[0041] (48 dimensions) and a migration covariance matrix [0042]
[External Character 8]
$W_x$

[0043] (48x48 dimensions) are searched for. These are called a perimeter pattern. Next, a standard pattern is created using the average and a covariance.
a. [0044] which communalizes a covariance matrix
[Equation 4]

$$W = \left( \overline{\sum_h \sum_j} W_{h,j} + g \cdot W_x \right) / ( 1 + g )$$

（$\overline{\sum \sum}$ は「平均をとる」の意味）

[0045] In the case of valve flow coefficient, h is about 130 by the class of voice piece here. Moreover, g is a rate which mixes a perimeter pattern and is usually g= 1. It carries out.
[0046] b. The partial pattern of each voice piece [0047]
[External Character 9]
$A_{h,j}$ 、 $B_{h,j}$

[0048] It creates.
[0049]
[Equation 5]

$$A_{h,j} = 2 ( \mu_{h,j} W^{-1} - \mu_x W^{-1} )$$

[0050]
[Equation 6]

$$B_{h,j} = \mu_{h,j} W^{-1} \mu_{h,j}{}^t - \mu_x{}^1 W^{-1} \mu_x{}^t$$

[0051] Derivation of these formulas is mentioned later. The example of the voice piece standard-pattern creating method is shown in drawing 4 . A frame response with a correlation sample is searched for between the start edges and termination of the sample for study, and it divides a voice piece sample into JS. It asks for a response frame with a representation sample in drawing 4 , and is (j). It is shown. And it is (j)-L1-(j)+L2 about each of (j) =1- (JS). The average and a covariance are calculated using the data for 100 pieces of the section, and it is a partial pattern [0052].
[External Character 10]
$A_{h,j}$、 $B_{h,j}$

[0053] ********. Therefore, voice piece h A standard pattern becomes what connected and gathered up the partial pattern of Jh individual including the section which overlaps **. A perimeter pattern asks for the average and a covariance, shifting the one L1+L2+1 frame partial section at a time, as shown in drawing. Not only the voice section but the noise section of order of the range of perimeter pattern creation is good also as an object. The voice piece standard pattern obtained about each word is beforehand stored in the voice piece standard-pattern storing section 6.
[0054] The link information of a voice piece is described by the [voice piece connection] word dictionary 5 for every word to recognize, and the example is shown in drawing 5 . The word standard-pattern generation section 7 reads and connects the voice piece standard pattern stored in the voice piece standard-pattern storing section 6 according to this voice piece link information. Of this connection actuation, as shown in the example of drawing 6 , the false standard pattern (it is hereafter described as a "word standard pattern") of a word is formed. It is the word standard pattern of the word "k" created as mentioned above [0055]
[Equation 7]
$$A_{k,j} = 2(\mu_{k,j} W^{-1} - \mu_x W^{-1})$$

[0056]
[Equation 8]
$$B_{k,j} = \mu_{k,j} W^{-1} \mu_{k,j}^{t} - \mu_x^{1} W^{-1} \mu_x^{t}$$

[0057] It expresses. In addition, in the case of drawing 2 , the data which performed inter-frame length in the 1st word standard-pattern processing section 8 are used as a word standard pattern as above-mentioned. In the case of drawing 3 , inter-frame length is not performed, but in the 2nd word standard-pattern processing section 13,

the standard pattern into which the environmental standard pattern stored in the ends of a word standard pattern at the environmental standard-pattern storing section 14 was added and processed is used.

[0058] The distance between the word standard pattern and two or more frame buffers which are [count of partial distance] above, and were made and formed (partial distance) is calculated in the partial distance count section 22. In addition, since it is collating about inter-frame length data in the case of drawing 2 , the suffixes i and j showing the frame number used by future explanation shall newly regive a number about the frame which performed inter-frame length.

[0059] Between input vectors and the partial patterns of each word including the information on the multiple frame shown by (several 1), count of partial distance is calculated using a statistical interval scale. Since the distance as the whole word will accumulate and find distance (partial distance) with a partial pattern, it needs to calculate partial distance irrespective of the location of an input, or the difference in a partial pattern by the approach which a distance value can compare mutually. For that, it is necessary to use the interval scale based on a-posteriori probability. Namely, the j-th partial pattern of an input (several 1) and the word "k" [0060]

[External Character 11]

$\omega_{k,\,j}$

[0061] About distance, it is a-posteriori probability [0062].

[External Character 12]

$P(\omega_{k,\,j} \mid X_i)$

[0063] Therefore, it calculates. It becomes like a degree type by Bayes' theorem.

[0064]

[Equation 9]

$$P(\omega_{k,\,j} \mid X_i) = P(\omega_{k,\,j}) \cdot P(X_i \mid \omega_{k,\,j}) / P(X)$$

[0065] The 1st term of the right-hand side considers that the appearance probability of each word is the same, and deals with it as a constant. The a-priori probability of the 2nd term of the right-hand side considers distribution of a parameter to be normal distribution, and becomes like a degree type.

[0066]

[Equation 10]

$$P(X_i \mid \omega_{k,j}) = (2\pi)^{-d/2} |W_{k,j}|^{-1/2} \cdot \exp\{-1/2 (X_i - \mu_{k,j}) W_{k,j}^{-1} (X_i - \mu_{k,j})^t\}$$

[0067] (Several 10) is the sum of a probability to all the input conditions that may occur also including a word and its circumference information, and a parameter can think that it becomes a distribution configuration near normal distribution in the case of an LPC cepstrum multiplier or a band pass filter output. Here, for (several 10), an average and a covariance are [0068], respectively.

[External Character 13]

$\mu_x$、 $W_x$

[0069] It is assumed that it is a thing according to **********.

[0070]

[Equation 11]

$$P(X) = (2\pi)^{-d/2} |W_x|^{-1/2} \cdot \exp\{-1/2 (X_i - \mu_x) W_x^{-1} (X_i - \mu_x)^t\}$$

[0071] A degree type will be obtained, if substitute (several 10) and (several 11) for (several 9), a logarithm is taken, a constant term is omitted and it doubles -two further.

[0072]

[Equation 12]

$$L_k(i,j) = (X_i - \mu_{k,j}) W_{k,j}^{-1} (X_i - \mu_{k,j})^t - (X_i - \mu_x) W_x^{-1} (X_i - \mu_x)^t + \log(|W_{k,j}| / |W_x|)$$

[0073] This formula is a formula which carried out a-posteriori probability of the BEIZU distance, and although discernment capacity is high, the fault that there is much computational complexity has it. This formula is developed to a linearity discriminant as follows. Covariance matrices also including all the partial patterns and perimeter patterns to all words assume that it is an equal. A covariance matrix is communalized by (several 4) on the basis of such an assumption, and if it substitutes for (several 12) and arranges, the easy following linearity discernment types can be drawn.

[0074]

[Equation 13]

$$L_k(i, j) = B_{k,j} - A_{k,j} \cdot X_i^t$$

$$A_{k,j} = 2(\mu_{k,j} W^{-1} - \mu_x W^{-1})$$

$$B_{k,j} = \mu_{k,j} W^{-1} \mu_{k,j}^t - \mu_x^1 W^{-1} \mu_x^t$$

[0075]

[External Character 14]

$A_{k,j}$、$B_{k,j}$

[0076] ** (several 7) and (several 8) will already show, and the j-th standard pattern of the word "k" will be expressed by this pair.

[0077] The distance accumulation section 23 is a part which accumulates to the section of partial distance j=1-Jk to each word, and asks for the similarity to the whole word. In that case, it is necessary to accumulate, expanding and contracting an input part (I frames) in the allowed-time length Jk of each word. This count is efficiently calculable using the technique (the DP method) of a dynamic programming.

[0078] In drawing 2 , since it is with the word-spotting method by collating the endpoint free-lancer who does not perform voice section detection, processing of word collating is as follows. Since voice section detection is not performed in the case of the word-spotting method, a collating operation is performed about sufficient long section containing voice. That is, using as the audio start edge i= 1 which it is at the collating initiation event like [ in the case of performing voice section detection ], and carrying out a collating operation for i=I as audio termination does not have semantics. By the word-spotting method, the collating score to a word standard pattern is computed by assuming an audio always edge about all the input sections. That is, the accumulation operation of partial similarity performed in the path judging 24 is as follows. Here, the subscript k of a word number is omitted, the partial distance of the i-th frame part of an input and the j-th partial pattern will be expressed as L (i, j), and the accumulation distance to a frame (i, j) will be expressed as g (i, j). The path judging section 24 is [0079].

[Equation 14]

$$g(i,1)=L(i,1)$$

$$g(i,j)=\min\begin{cases}g(i-2,j-1)+L(i,j)\\ g(i-1,j-1)+L(i,j)\\ g(i-1,j-2)+L(i,j-1)+L(i,j)\end{cases}$$

$$(1<i\leqq I,\ 1<j\leqq J)$$

[0080] ****** is performed and the path from which accumulation distance becomes min among three paths shown by the formula is chosen. Thus, in the judgment section 25, use this g (i, J) as the final collating score of a word standard pattern, and after accumulating distance serially, when g (i, J) takes the smallest value to i, let i at this time be audio termination. The audio start edge can be obtained by following the path which the path judging section 24 judged.

[0081] Since the operation performed in the path judging section 34 of drawing 3 is processing of endpoint immobilization, it is as follows.

[0082]

[Equation 15]

$$g(1,1)=L(1,1)$$

$$g(i,j)=\min\begin{cases}g(i-2,j-1)+L(i,j)\\ g(i-1,j-1)+L(i,j)\\ g(i-1,j-2)+L(i,j-1)+L(i,j)\end{cases}$$

$$(1<i\leqq I,\ 1<j\leqq J)$$

[0083] In (several 15), for convenience, the audio frame i is reattached so that the start edge of the collating section may be set to 1 and termination may be set to I in a number.

[0084] The path judging section 34 chooses the path from which accumulation distance becomes min among three paths shown by (several 15). Thus, distance is accumulated serially and it considers as the collating score of the accumulation

distance g (i, Jk) word in the event of becoming j=Jk and i=I "k." The judgment section 35 is outputted in quest of the word "k" which gives the minimum value of the accumulation distance g (i, Jk).

[0085] Hereafter, the 2nd example of this invention is explained. Drawing 7 shows the block diagram of the 2nd example of this invention. In drawing 7 , the same number is given to the same component as drawing 1 . Different points from the 1st example are 1st collating/judgment section 4 and the word standard-pattern merge section 41, and explain in detail the content of processing performed by these next. First, the content of processing performed in the word standard-pattern merge section 41 is explained. For the computational complexity cutback by partial distance count, two partial standard patterns are made into a group, and are packed into one. Since the linearity discriminant is used, it is equal to finding partial distance to ask for the sum of the partial distance on DP pass, after adding a corresponding parameter previously. Therefore, this processing will fix DP pass in every two frames to one. Partial standard pattern [0086]

[External Character 15]

$\widehat{A}_{k,n}$

[0087] [0088] of ******

[External Character 16]

$A_{k,2n-1}$

[0089] [0090]

[External Character 17]

$A_{k,2n}$

[0091] It shifts by one frame, merges and creates. That is, it is [0092] when the conventional partial distance is shown in (several 16) and (several 17).

[Equation 16]

$$L_k(i, 2n-1) = B_{k,2n-1} - A_{k,2n-1} \cdot X_{i-1}^{t}$$

[0093]

[Equation 17]

$$L_k(i, 2n) = B_{k,2n} - A_{k,2n} \cdot X_i^{t}$$

[0094] Partial distance summarizes the two above-mentioned formula, and is [0095].

[Equation 18]

$$\widehat{L}_k(i,n) = \widehat{B}_{k,n} - \widehat{A}_{k,n} \cdot \widehat{X}_i^{\,t}$$

[0096] It becomes equal to finding the conventional partial distance a condition [ fixing a next door and DP pass in every two frames to one ].

[0097] If the feature parameter used for partial distance count is L frames by this amelioration (L+1), computational complexity is reducible to /2L. In the 2nd example, as a parameter if L= 4, the computational complexity in this case is reducible to five eighths.

[0098] Next, the content of processing performed in 1st collating/judgment section 42 is explained using a drawing. Drawing 8 is the block diagram showing the detail of the flow of processing of 1st collating/judgment section 42. The partial distance of the word standard pattern and two or more frame buffers which are obtained from the 1st word standard-pattern processing section 8 and the word standard-pattern merge section 41 is calculated in the partial distance count section 22. It is representation partial distance [0099] beforehand to the partial distance for two frames to an input about the one where distance is smaller at the representation partial distance selection section 51.

[External Character 18]

$\mathrm{typ}\,\widehat{L}_k(2i,n)$

[0100] It will be set to (several 19) if it carries out.

[0101]

[Equation 19]

$$\mathrm{typ}\,\widehat{L}_k(2i,n) = \min\left(\widehat{L}_k(2i-1,n), \widehat{L}_k(2i,n)\right)$$

[0102] About this representation partial distance, it accumulates in the distance accumulation section 23, and asks for the similarity to the whole word. In that case, it is necessary to accumulate, expanding and contracting an input part (I frames) in the allowed-time length Jk of each word. This count is efficiently calculable using the DP method like the 1st example.

[0103] In drawing 8 , since it is with the word-spotting method by collating the endpoint free-lancer who does not perform voice section detection, processing of word collating is as follows. Since voice section detection is not performed in the case of the word-spotting method, a collating operation is performed about sufficient long section containing voice. By the word-spotting method, the collating score to a word

standard pattern is computed by assuming an audio always edge about all the input sections. That is, the accumulation operation of partial similarity performed in the path judging 52 is as follows. The subscript k of a word number is omitted, the partial distance of the i-th frame part of an input and the j-th partial pattern is expressed as typL (i, j) here, and it is the accumulation distance to a frame (i, j) [0104]
[External Character 19]
$\widehat{g}(i,j)$

[0105] It will express. The path judging section 52 is [0106].
[Equation 20]

$$\widehat{g}(i,1) = \mathrm{typ}\widehat{L}(i,1)$$

$$\widehat{g}(i,j) = \min \begin{cases} \widehat{g}(i-1,j-1) + \mathrm{typ}\widehat{L}(i,j) \\ \widehat{g}(i-2,j-1) + \mathrm{typ}\widehat{L}(i,j) \\ \widehat{g}(i-3,j-1) + \mathrm{typ}\widehat{L}(i,j) \end{cases}$$

$$(1 < i \leq I,\ 1 < j \leq J)$$

[0107] ****** is performed and the path from which accumulation distance becomes min among three paths shown by the formula is chosen. thus, the time of g (i, J) taking the smallest value to i in the judgment section 25, after accumulating distance serially -- this -- Use g (i, J) as the final collating score of a word standard pattern, and let i at this time be audio termination. The audio start edge can be obtained by following the path which the path judging section 52 judged. Henceforth, the same processing as the 1st example is performed.
[0108]
[Effect of the Invention] As explained above, after this invention narrows down a word candidate beforehand by the collating operation using the data thinned out first If it is effective in aiming at the cutback of computational complexity and the case where it realizes by the hardware of the same magnitude is considered, securing the recognition engine performance by performing collating using the data which do not cull out to the narrowed-down candidate The number of vocabularies is expandable about single figure, maintaining the recognition engine performance compared with the conventional example. Moreover, in order not to perform detection of that the phenomenon of partial matching which incorrect recognition produces when the word

which poses a problem collates with the one section of other words by collating including the section of the outside of the voice section by using a word standard pattern with an environmental pattern in the case of word spotting does not arise, and the precise voice section, either, it becomes realizable [ a dogged voice recognition unit ] to a noise.

[0109] Furthermore, the count of a comparison operation is substantially reduced by count of a partial product reducing and reducing the lattice points of DP to each of a dictionary shaft and input shafts 1/2 by merging a partial standard pattern for every multiple frame in the 2nd example.

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平6－266393

(43)公開日　平成６年(1994)９月22日

| (51)Int.Cl.[5] | | 識別記号 | 庁内整理番号 | ＦＩ | 技術表示箇所 |
|---|---|---|---|---|---|
| Ｇ１０Ｌ | 5/06 | | Ｆ　9379－5H | | |
| | 3/00 | ５２１ | Ｌ　9379－5H | | |
| | | | Ｎ　9379－5H | | |
| | | ５３１ | Ｆ　9379－5H | | |
| | | | Ｃ　9379－5H | | |

審査請求　未請求　請求項の数５　ＯＬ　（全 16 頁）　最終頁に続く

(21)出願番号　特願平5－51897

(22)出願日　平成５年(1993)３月12日

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　木村　達也
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内

(72)発明者　▲桑▼野　裕康
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内

(72)発明者　渡辺　泰助
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内

(74)代理人　弁理士　小鍛冶　明　（外２名）

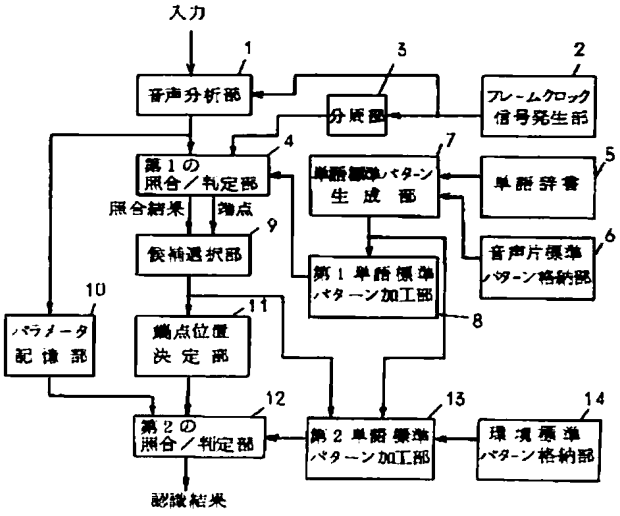
最終頁に続く

(54)【発明の名称】　音声認識装置

(57)【要約】

【目的】　本発明は不特定話者を対象とした音声の自動認識に関し、計算量が少なくノイズに強い高精度の大語彙認識が可能な音声認識装置を提供することを目的とする。

【構成】　全候補単語に対して、ワードスポッティングの機能を有する計算量の少ない大まかな第一の照合により予備選択を行い候補単語を絞り込んだ後、絞り込まれた候補に対してのみ精密な第二の照合を行うという２段階の処理により認識を行うことにより、大語彙認識を少ない計算量で実現する。また、上記予備選択により音声区間を確実に含む区間を決定し、単語標準パターンの前後に環境標準パターンを接続することによって得られる標準パターンを用いて上記第二の照合を行うことにより、音声区間の前後のノイズに対して強い認識を実現する。



1

【特許請求の範囲】

【請求項１】　入力音声信号を分析の基本単位であるフレーム毎に分析し分析パラメータを抽出する音声分析手段と、分析フレームのタイミング信号を発声するフレームクロック発生手段と、上記フレームクロックを所定の分周比で分周して分周クロック信号を得る分周手段と、単語を音声片を表す記号の系列で表記した単語辞書と、予め多数の人が発声した音声データを用いて作成される声片データの一部を表現する部分標準パターンの系列により構成される音声片標準パターンを格納する音声片標準パターン格納手段と、上記音声片標準パターンを上記単語辞書の表記内容に従って接続することにより単語の標準パターンを得る単語標準パターン生成手段と、上記単語標準パターンを構成する特徴パラメータ系列の一部を間引くことによりデータ間引き単語標準パターンを作成する第１の単語標準パターン加工手段と、上記分周クロック信号を受け取る毎に同時点における上記分析パラメータから得られる特徴パラメータと上記データ間引き単語標準パターンの部分標準パターンとの間の距離である部分距離を算出し、当該時点およびそれ以前の特徴パラメータ系列に対して既に求められている上記データ間引き単語標準パターンとの間の部分距離を累積することにより、当該時点を単語の終端と仮定した場合のデータ間引き単語標準パターンの入力に対する最小の距離およびそれに付随する始端位置を得て、上記分周クロック毎に上記最小の距離を上記始端位置と併せて各単語毎に更新する第１の照合／判定手段と、入力音声の終了時点で全認識対象単語の標準パターンに対する距離を相互に比較することにより距離値の小さい順に所定の個数候補単語を得る候補単語選択手段と、上記候補選択手段によって選択された候補単語に付随する始端および終端候補群から音声区間を確実に含む区間を決定する端点位置決定手段と、上記分析パラメータを全入力区間にわたって記憶するパラメータ記憶手段と、あらかじめ音声以外の区間の音響信号から作成された環境標準パターンを格納する環境パターン格納手段と、上記環境パターンを上記単語標準パターンの前後に接続して環境パターンつき単語標準パターンを作成する第２の単語標準パターン加工手段と、上記単語候補選択手段により選択された単語候補群に対応する環境パターンつき単語標準パターンと上記パラメータ記憶手段に格納されている上記端点位置決定手段によって決定された区間におけるパラメータ系列との間の距離を部分距離の累積することにより算出し、上記候補単語毎に得られる距離を相互比較することによりもっとも値の小さい距離値を得た単語候補を認識結果として出力する第２の照合／判定手段とからなる音声認識装置。

【請求項２】　第１の照合／判定手段の処理においてフレーム間引きを利用して部分距離の算出と単語照合における計算を簡略化することを特徴とする請求項１記載の音声認識装置。

2

【請求項３】　第１の単語標準パターン加工手段で作成されたデータ間引き単語標準パターンの部分標準パターンを同一複数フレームを組として１つにまとめることによりデータ間引き併合単語標準パターンを作成する第１の単語標準パターン併合手段を付加し、第１の照合／判定手段は、分周クロック信号を受け取る毎に同時点における分析パラメータから得られる特徴パラメータと上記データ間引き併合単語標準パターンの部分標準パターンとの間の距離である部分距離を算出する部分距離計算部と、前記部分距離と当該時点より前時点の部分距離を比較し距離の小さい方を代表部分距離とする代表部分距離選択部と、当該時点およびそれ以前の特徴パラメータ系列に対して既に求められている上記データ間引き単語標準パターンとの間の代表部分距離を累積する距離累積部と、当該時点を単語の終端と仮定した場合のデータ間引き単語標準パターンの入力に対する最小の距離およびそれに付随する始端位置を得て、上記分周クロック毎に上記最小の距離を上記始端位置と併せて各単語毎に更新する判定部とを有し、部分距離の算出および単語照合における計算を簡略化することを特徴とする請求項１記載の音声認識装置。

【請求項４】　部分距離は統計的距離尺度を用いて算出し、上記統計的距離尺度が事後確率に基づく距離尺度であることを特徴とする請求項１乃至３のいずれかに記載の音声認識装置。

【請求項５】　部分距離は統計的距離尺度を用いて算出し、上記統計的距離尺度が事後確率に基づく一次判別式であることを特徴とする請求項１乃至３のいずれかに記載の音声認識装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は人間の声を機械に認識させる音声認識の方法に関するものである。

【０００２】

【従来の技術】音声認識には特定話者を対象とする方法と、不特定話者を対象とする方法とがあるが、本発明は特に不特定話者認識を対象とするものである。不特定話者を対象とした方法の一例として、特願平３－３１４２４８号に基づく例を図９を参照しながら説明する。

【０００３】図７において、61は音響分析部、62は特徴パラメータ抽出部、63は音声区間検出部、64は複数フレームバッファ、65は音声片標準パターン格納部、66は全認識対象単語を音声片の並びで記述した単語辞書、67は音声片の並びに従って音声片標準パターンを選択し連結することにより認識対象語彙の単語標準パターンを生成する単語標準パターン生成部、68は複数のフレームで形成された入力ベクトルと認識対象音声の部分パターンとの部分距離を事後確率に基づく統計的距離尺度で求める部分距離計算部、69は入力フレームをシフトしながら音

声全体にわたって部分距離を累積することにより入力音声と単語標準パターンとの距離を求める距離累積部、610は経路判定部、611は累積距離を最小とする音声名を認識結果とする判定部である。

【０００４】音響分析部61は入力信号をＡＤ変換して一定時間長（フレームと呼ぶ。本従来例では10ms）毎に分析する。特徴パラメータ抽出部62では音響分析部61の出力結果に基づき、特徴パラメータを抽出する。音声区間検出部63は入力信号音声の始端、終端を検出する。音声区間の検出法は音声のパワーを用いる方法が簡単で一般的であるが、どのような方法でもよい。また、音声区間検出を行わずに、入力の全区間を端点と仮定して照合演算を行うワードスポッティングという方法を用いる場合もあり、これについては後述する。複数フレームバッファ64は各フレームの近隣のフレームの特徴パラメータを統合してパターンマッチング（部分マッチング）に用いる入力ベクトルを形成する部分である。音声片標準パターン格納部65には音声片の標準パターンを部分パターンの結合として格納しておく。単語辞書66には認識する単語毎に音声片の連結情報が記述されている。音声片連結部67はこの音声片連結情報に従って音声片標準パターン格納部65に格納されている音声片標準パターンを読み出し連結する。部分距離計算部68において単語標準パターンと複数フレームバッファとの間の距離（部分距離）を計算する。距離累積部69は、各単語に対する部分距離を累積し、単語全体に対する類似度を求める。経路判定部610は累積距離が最小になる経路を選択する。判定部611は、累積距離の最小値を与える単語を求め出力する。

【０００５】次に、音声区間検出を行わないワードスポッティング法を用いる場合について説明する。ワードスポッティング法の利点は、一般にノイズに弱いとされる音声区間検出を用いる必要がないため、ノイズに強い認識系が実現できることである。ワードスポッティング法の場合、音声区間検出を行わないので、照合演算は音声を含む十分長い区間について行う。すなわち、音声区間検出を行う場合の様に、照合開始時点を音声の始端とし、照合終了時点を音声の終端として照合演算をすることは意味を持たない。ワードスポッティング法では、全入力区間について音声の始終端を仮定して単語標準パターンに対する照合スコアを算出する。

【０００６】

【発明が解決しようとする課題】従来例で説明した方法は「近隣フレーム間の時間的な動き」の情報を積極的に利用している事と、統計的距離尺度を用いる事により精度の高い不特定話者用音声認識が可能である。また、音声片を連結する方法であるので、単語辞書を書換えるだけで語彙変更可能な汎用性の高い認識装置の実現が可能である。更にワードスポッティングを行うことにより精密な音声区間検出が不要となるため、ノイズに強い認識装置を実現できる。

【０００７】しかしこの方法は、特徴パラメータとしてあるフレームの近隣フレームを含めた幅のある区間（複数フレーム）の分析パラメータを用いているために特徴パラメータの次元数が大きいことに加え、部分距離を全入力区間及び、単語の標準パターンの全区間について求めているために、部分距離の計算に線形識別式を用いているとはいえ、まだ計算量が多いという問題点があった。また、ワードスポッティングを用いた場合に、「藤井寺」と「富士」の例のように、ある単語が他の単語の一部分とマッチングして誤認識となる「部分マッチング」の問題があった。

【０００８】

【課題を解決するための手段】以上述べた問題を解決するために本発明では、入力音声信号を分析の基本単位であるフレーム毎に分析し分析パラメータを抽出する音声分析手段と、分析フレームのタイミング信号を発声するフレームクロック発生手段と、上記フレームクロックを所定の分周比で分周して分周クロック信号を得る分周手段と、単語を音声片を表す記号の系列で表記した単語辞書と、予め多数の人が発声した音声データを用いて作成される音声片データの一部を表現する部分標準パターンの系列により構成される音声片標準パターンを格納する音声片標準パターン格納手段と、上記音声片標準パターンを上記単語辞書の表記内容に従って接続することにより単語の標準パターンを得る単語標準パターン生成手段と、上記単語標準パターンを構成する特徴パラメータ系列の一部を間引くことによりデータ間引き単語標準パターンを作成する第１の単語標準パターン加工手段と、上記分周クロック信号を受け取る毎に同時点における上記分析パラメータから得られる特徴パラメータと上記データ間引き単語標準パターンの部分標準パターンとの間の距離である部分距離を算出し、当該時点およびそれ以前の特徴パラメータ系列に対して既に求められている上記データ間引き単語標準パターンとの間の部分距離を累積することにより、当該時点を単語の終端と仮定した場合のデータ間引き単語標準パターンの入力に対する最小の距離およびそれに付随する始端位置を得て、上記分周クロック毎に上記最小の距離を上記始端位置と併せて各単語毎に更新する第１の照合／判定手段と、入力音声の終了時点で全認識対象単語の標準パターンに対する距離を相互に比較することにより距離値の小さい順に所定の個数候補単語を得る候補単語選択手段と、上記候補選択手段によって選択された候補単語に付随する始端および終端候補群から音声区間を確実に含む区間を決定する端点位置決定手段と、上記分析パラメータを全入力区間にわたって記憶するパラメータ記憶手段と、あらかじめ音声以外の区間の音響信号から作成された環境標準パターンを格納する環境パターン格納手段と、上記環境パターンを上記単語標準パターンの前後に接続して環境パターンつき単語標準パターンを作成する第２の単語標準パター

ン加工手段と、上記単語候補選択手段により選択された単語候補群に対応する環境パターンつき単語標準パターンと上記パラメータ記憶手段に格納されている上記端点位置決定手段によって決定された区間におけるパラメータ系列との間の距離を部分距離の累積することにより算出し、上記候補単語毎に得られる距離を相互比較することによりもっとも値の小さい距離値を得た単語候補を認識結果として出力する第２の照合／判定手段とを設ける。

【０００９】

【作用】本発明は、上記各手段とりわけ分周手段と第１の単語標準パターン加工手段によって間引かれたデータを用いて照合演算によってあらかじめ単語候補を絞り込んだ後に、絞り込まれた候補に対して間引きをしないデータを用いた照合を行うことにより認識性能を確保しながら計算量の削減を図るという第１の作用効果と、上記環境パターンつき単語標準パターンを用いることにより音声区間の外側の区間を含めて照合をすることにより、ノイズに強い認識を実現するとともに、ワードスポッティングの場合に問題となっていたある単語が他の単語の１部と照合することにより誤認識が生ずる部分マッチングの問題を解決するという第２の作用効果をもつものである。

【００１０】

【実施例】以下、図面を用いて本発明の第１の実施例を説明する。図１は本発明の第１の実施例の構成を示したものである。図１において、１は音響分析部、２はフレームクロック信号発生部、３は分周部、４は第１の照合／判定部、５は単語辞書、６は音声片標準パターン格納部、７は単語標準パターン生成部、８は第１単語標準パターン加工部、９は候補選択部、１０はパラメータ記憶部、１１は端点位置決定部、１２は第２の照合／判定部、１３は第２単語標準パターン加工部、１４は環境標準パターン格納部である。次にその動作を説明する。

【００１１】音響分析部１は入力信号をＡＤ変換して一定時間長（フレームと呼ぶ。本実施例では10ms）毎に分析する。例では線形予測分析（ＬＰＣ分析）を用いている。フレームのタイミングはフレームクロック信号発生部２が発生するクロック信号により与えられ、このクロック信号は音響分析部１および分周部３に供給される。分周部３はフレームクロック信号を所定の分周比（本実施例では２）で分周して分周クロック信号を出力する。この分周クロック信号は第１の照合/判定部４へ供給されフレーム間引きのために使用される。第１の照合/判定部４は音響分析部１の出力する分析パラメータと後述の処理によって生成される単語標準パターンとの間の照合をワードスポッティングにより行う。この部分の処理の詳細については後述する。

【００１２】次に、第１の照合／判定部４で使用する単語標準パターンの生成方法について、説明する。認識語彙はＣＶやＶＣ等の音声片記号の並びで表現され、単語辞書５に格納されている。単語標準パターンは、単語辞書５を参照して得られる音声片記号の並びに従って、音声片標準パターン格納部６に格納されている音声片標準パターンを連結することにより単語標準パターン生成部７によって生成される。

【００１３】なお音声片標準パターンの作成方法については後述する。第１の照合／判定部４における照合は、計算量削減のため、フレーム間引きがなされたデータに対して行われる。従って第１の照合／判定部４で用いる単語標準パターンもフレーム間引きを施す必要がある。第１単語標準パターン加工部８はこのフレーム間引きの処理を、単語標準パターン生成部７で得られた単語標準パターンに対して行う。候補選択部９は、第１の照合／判定部４で得られた全ての単語に対する照合結果から、照合結果の良い順に所定の個数だけ単語候補を選択する。パラメータ記憶部１０は音声分析部１で得られた分析パラメータを全入力区間について記憶する。

【００１４】端点位置決定部１１では、上記単語候補の各々に付随して得られる始端及び終端の情報を統合して、第２の照合／判定部１２で照合を行うための照合区間を決定する。この照合区間は音声区間を必ず含むように決定される。従って実際の音声区間より長い区間が得られる。例えば、単語候補に付随して得られる始端群の先頭もしくは更に前の位置が照合区間の始端として決定される。終端の場合も同様であり、単語候補に付随して得られる終端群の末尾もしくはさらに後の位置が照合区間の終端として決定される。

【００１５】第２の照合／判定部１２ではパラメータ記憶部１０に記憶されたパラメータと後述の処理に従って得られる単語標準パターンに対して、上記単語候補および照合区間について、間引きを行わない端点固定の照合を行ったのち、最も良い照合結果を与える単語候補を認識結果として出力する。第２の照合／判定部１２に与えられる照合区間は上述の通り、実際の音声区間より長くとられるため、照合に用いる単語標準パターンは、第２単語標準パターン加工部１３によって、単語標準パターン生成部７により得られた単語標準パターンの両端に環境標準パターン格納部１４に格納されている環境標準パターンを接続する処理を施したものを使用する。この環境標準パターンは例えば、あらかじめ認識装置が使用される騒音信号のパターンから作成される。

【００１６】次に第１の照合/判定部４及び第２の照合／判定部１２で行う処理内容について詳しく説明する。両者の相違は、前者はフレーム間引きをしたデータについて照合演算をしているのに対し後者では間引きをしないデータについて照合演算を行っていることと、前者は照合の区間を与えない端点フリーの照合によるワードスポッティングを行っているのに対し、後者は照合区間をあらかじめ与える端点固定の照合を行っている点であ

る。その他の、照合に用いる特徴パラメータや使用する距離尺度等の基本的な考え方は同じであるので、照合処理の詳細は、第１の照合/判定部４について述べ相違がある部分についてはその都度説明をする。

【００１７】図２は第１の照合/判定部４の処理の流れを示す詳細な構成図、図３は第２の照合/判定部１２の構成図である。説明は主に図２を用いて行うが、必要に応じて図３及び、図３を参照する。また、図２及び図３の各構成要素について、名称が同じものは同じ機能を有する。図２及び図３において、２１は複数フレームバッファ、２２は部分距離計算部、２３は距離累積部、２４は経路判定部、２５は判定部である。



【００１８】図２において、複数フレームバッファ２１は第ｉフレームの近隣のフレームの特徴パラメータを統合してパターンマッチング（部分マッチング）に用いる入力ベクトルを形成する部分である。第ｉフレームにおける入力ベクトル

【００１９】

【外１】

$$X_i$$

【００２０】は、次のように表わされる。

【００２１】

【数１】

$$X_i = (x_{i-L1},\ x_{i-L1+m},\ \cdots,\ x_i,\ \cdots,\ x_{i+L2})$$

【００２２】これはｍフレームおきにｉ－Ｌ１～ｉ＋Ｌ２フレームの特徴パラメータを統合したベクトルである。Ｌ１=Ｌ２=３，ｍ=２ とするとＸｉの次元数は$(p+2)\times\{(L1+L2+1)/m+1\}=12\times4=48$となる。ｍが２以上の値をとる場合にはフレームを間引いて入力ベクトルを形成することに相当する。音声片標準パターン格納部６は音声片の標準パターンを部分パターンの結合として格納してある部分である。ここで音声片標準パターン作成法をやや詳細に説明する。

【００２３】［音声片標準パターン作成方法］音声片とは、音声認識の基本単位として用いる音声の素片であり、種類としては音素、音節（ＣＶ）、半音節（ＶＣ、ＣＶ）、母音－子音－母音連鎖（ＶＣＶ）等がある。なおＣは子音をＶは母音を意味する。以下の説明では、一例として音声片の種類として音節（ＣＶ）を用いる場合について説明を行う。

【００２４】例えば音声片/sa/の標準パターンは次のような手段で作成する。

$$S_j = (s_{j-L1},\ s_{j-L1+m},\ \cdots,\ s_j,\ \cdots,\ s_{j+L2})\ (j=1,\cdots,JS)$$

（１）多数の人が発声した音声データから、/sa/と発声している部分を切り出す（１００個サンプルが切り出されているとする）。

（２）１００個の/sa/の持続時間分布を調べ、１００個の平均時間長JSを求める。

（３）JSの時間長のサンプルを１００個の中から探し出す。複数のサンプルがあった場合はフレームごとに複数サンプルの平均値を計算する。このように求められた代表サンプル

【００２５】

【外２】

$$S_j$$

【００２６】を

【００２７】

【数２】

【００２８】とする。ここでｓｊは１フレームあたりのパラメータベクトルであり、分析パラメータと同様に１１個のＬＰＣケプストラム係数と差分パワーで構成される。

（４）１００個分の各サンプル（数１）と代表サンプル（数２）との間でパターンマッチングを行ない、代表サンプルのフレームと１００個分の各サンプルのフレーム間の対応関係を求める（最も類似したフレーム同志を対応づける）。なお、フレーム間の対応関係は例えばダイナミックプログラミングの手法を用いれば効率よく求めることができる。

（５）代表サンプルの各フレーム（ｊ=１～JS）に対応して、１００個分のサンプルそれぞれから（数１）の形の部分ベクトルを切り出す。簡単のため l1＝l2＝3、ｍ＝１とする。

【００２９】代表サンプルの第ｊフレームに相当する、１００個分のデータのうち第ｎ番目のサンプルの部分ベクトルを

【００３０】

【数３】

$$X^n_j = (x^n_{j-L1},\ x^n_{j-L1+m},\ \cdots,\ x^n_j,\ \cdots,\ x^n_{j+L2})$$

【００３１】とする。ここでｊは同一単語/sa/の第ｎ番目のサンプル中、代表ベクトルの第ｊフレームに対応するフレームであることを示す。本実施例では４８次元のベクトルである（ｎ＝１～100）。

（６）１００個の

【００３２】

【外３】

$$X^n_j$$

【００３３】の平均値

9

【００３４】
【外４】
$\mu_j$
【００３５】（４８次元）と共分散行列
【００３６】
【外５】
$W_j$
【００３７】（４８×４８次元）を求める（j＝1～JS）。平均値と共分散行列は標準フレーム長の数JSだけ存在することになる（ただし、これらは必ずしも全フレームに対して作成する必要はない。間引いて作成してもよい）。上記（１）～（６）同様の手続きで音声片/sa/以外の音声片に対しても
【００３８】
【外６】
$\mu_j$ 、 $W_j$
$$W=(\overline{\sum_h \sum_j} W_{h,j} + g \cdot W_x)/(1+g)$$
（$\overline{\sum \sum}$ は「平均をとる」の意味）
【００４５】ここでhは音声片の種類でＣＶの場合、１３０程度である。また、gは周囲パターンを混入する割合であり通常g＝1 とする。
【００４６】ｂ．各音声片の部分パターン
【００４７】
【外９】
$A_{h,j}$ 、 $B_{h,j}$
$$B_{h,j}=\mu_{h,j}W^{-1}\mu_{h,j}^{t}-\mu_x^{1}W^{-1}\mu_x^{t}$$
【００５１】これらの式の導出は後述する。音声片標準パターン作成法の例を図４に示す。学習用サンプルの始端と終端の間において、標準サンプルとのフレーム対応を求めて、それによって音声片サンプルをJSに分割する。図４では、代表サンプルとの対応フレームを求めて(j) で示してある。そして、(j)＝１～(JS)の各々について、(j)－L1～(j)＋L2 の区間の１００個分のデータを用いて平均値と共分散を計算し、部分パターン
【００５２】
【外１０】
$A_{h,j}$ 、 $B_{h,j}$
【００５３】を求める。従って、音声片h の標準パターンは互にオーバーラップする区間を含むJh個の部分パターンを接続して寄せ集めたものになる。周囲パターンは図のようにL1+L2+1フレームの部分区間を1フレームずつシフトさせながら平均値と共分散を求める。周囲パターン作成の範囲は音声区間のみならず前後のノイズ区間も対象としてもよい。各単語について得られた音声片標準

10

【００３９】を求める。全ての音声区間に対する全てのサンプルデータに対し、移動平均
【００４０】
【外７】
$\mu_x$
【００４１】（４８次元）と移動共分散行列
【００４２】
【外８】
$W_x$
【００４３】（４８×４８次元）を求める。これらを周囲パターンと呼ぶ。次に平均値と共分散を用いて標準パターンを作成する。
ａ．共分散行列を共通化する
【００４４】
【数４】
$$W=(\overline{\sum_h \sum_j} W_{h,j} + g \cdot W_x)/(1+g)$$
（$\overline{\sum \sum}$ は「平均をとる」の意味）
【００４８】を作成する。
【００４９】
【数５】
$$A_{h,j}=2(\mu_{h,j}W^{-1}-\mu_x W^{-1})$$
【００５０】
【数６】
$$B_{h,j}=\mu_{h,j}W^{-1}\mu_{h,j}^{t}-\mu_x^{1}W^{-1}\mu_x^{t}$$
パターンは音声片標準パターン格納部６にあらかじめ格納しておく。
【００５４】［音声片連結］単語辞書５には認識する単語毎に音声片の連結情報が記述され、図５にその例を示す。単語標準パターン生成部７はこの音声片連結情報に従って音声片標準パターン格納部６に格納されている音声片標準パターンを読み出し連結する。この連結操作により、図６の例に示すように単語の疑似的な標準パターン（以下、「単語標準パターン」と記す）が形成される。以上の様にして作成された単語ｋの単語標準パターンを
【００５５】
【数７】
$$A_{k,j}=2(\mu_{k,j}W^{-1}-\mu_x W^{-1})$$
【００５６】
【数８】
$$B_{k,j}=\mu_{k,j}W^{-1}\mu_{k,j}^{t}-\mu_x^{1}W^{-1}\mu_x^{t}$$
【００５７】と表わす。なお、前述の通り、図２の場合には、第１単語標準パターン加工部８でフレーム間引きを行ったデータを単語標準パターンとして用いる。図３の場合にはフレーム間引きは行わないが第２単語標準パ

11

ターン加工部１３において、単語標準パターンの両端に環境標準パターン格納部１４に格納されている環境標準パターンを付加して加工した標準パターンを用いる。

【００５８】［部分距離の計算］上記のようにして形成された単語標準パターンと複数フレームバッファとの間の距離（部分距離）を部分距離計算部２２において計算する。なお、図２の場合にはフレーム間引きデータについて照合を行っているので今後の説明で用いるフレーム番号を現わす添え字ｉおよびｊはフレーム間引きを行ったフレームについて新たに番号をつけ直すものとする。

【００５９】部分距離の計算は(数１)で示す複数フレームの情報を含む入力ベクトルと各単語の部分パターンとの間で、統計的な距離尺度を用いて計算する。単語全体としての距離は部分パターンとの距離（部分距離）を累積して求めることになるので、入力の位置や部分パター

12

ンの違いにかかわらず距離値が相互に比較できる方法で部分距離を計算する必要がある。このためには、事後確率に基づく距離尺度を用いる必要がある。すなわち、入力(数１)と単語ｋの第j番目の部分パターン

【００６０】

【外１１】

$$\omega_{k,j}$$

【００６１】との距離を、事後確率

【００６２】

【外１２】

$$P(\omega_{k,j} \mid X_i)$$

【００６３】よって計算する。ベイズの定理により次式のようになる。

【００６４】

【数９】

$$P(\omega_{k,j} \mid X_i) = P(\omega_{k,j}) \cdot P(X_i \mid \omega_{k,j}) / P(X)$$

【００６５】右辺第１項は、各単語の出現確率を同じと考え、定数として取扱う。右辺第２項の事前確率は、パラメータの分布を正規分布と考え、次式のようになる。

【００６６】

【数１０】

$$P(X_i \mid \omega_{k,j}) = (2\pi)^{-d/2} |W_{k,j}|^{-1/2} \cdot \exp\{-1/2 (X_i - \mu_{k,j}) W_{k,j}^{-1} (X_i - \mu_{k,j})^t\}$$

【００６７】(数１０)は単語とその周辺情報も含めて、生起し得る全ての入力条件に対する確率の和であり、パラメータがＬＰＣケプストラム係数やバンドパスフィルタ出力の場合は、正規分布に近い分布形状になると考えることができる。ここでは（数１０）が、平均と共分散がそれぞれ

【００６８】

【外１３】

$$\mu_x,\ W_x$$

【００６９】の正規分布に従うものと仮定する。

【００７０】

【数１１】

$$P(X) = (2\pi)^{-d/2} |W_x|^{-1/2} \cdot \exp\{-1/2 (X_i - \mu_x) W_x^{-1} (X_i - \mu_x)^t\}$$

【００７１】（数１０）、（数１１）を（数９）に代入し、対数をとって、定数項を省略し、さらに－２倍すると、次式を得る。

【００７２】

【数１２】

$$L_k(i,j) = (X_i - \mu_{k,j}) W_{k,j}^{-1} (X_i - \mu_{k,j})^t - (X_i - \mu_x) W_x^{-1} (X_i - \mu_x)^t + \log(|W_{k,j}| / |W_x|)$$

【００７３】この式は、ベイズ距離を事後確率した式であり、識別能力は高いが計算量が多いという欠点がある。この式を次のようにして線形判別式に展開する。全ての単語に対する全ての部分パターンそして周囲パターンも含めて共分散行列が等しいものと仮定する。このような仮定のもとに共分散行列を（数４）によって共通化し、（数１２）に代入し整理すると次の様な簡単な線形識別式を導くことができる。

【００７４】

【数１３】

13　　14

$$L_k(i,j)=B_{k,j}-A_{k,j}\cdot X_i^{t}$$

$$A_{k,j}=2(\mu_{k,j}W^{-1}-\mu_x W^{-1})$$

$$B_{k,j}=\mu_{k,j}W^{-1}\mu_{k,j}^{t}-\mu_x^{1}W^{-1}\mu_x^{t}$$

【００７５】
【外１４】

$A_{k,j}$、$B_{k,j}$

【００７６】は（数７）、（数８）で既に示したものであり、この対で単語ｋの第j番目の標準パターンを表現していることになる。

【００７７】距離累積部２３は、各単語に対する部分距離j＝1～Jkの区間に対して累積し、単語全体に対する類似度を求める部分である。その場合入力部分（Ｉフレーム）を各単語の標準時間長Jkに伸縮しながら累積する必要がある。この計算はダイナミックプログラミングの手法（ＤＰ法）を用いて効率よく計算できる。

【００７８】図２では音声区間検出を行わない端点フリーの照合を行うことによりワードスポッティング法をもちいているので単語照合の処理は以下の様になる。ワードスポッティング法の場合、音声区間検出を行わないので、照合演算は音声を含む十分長い区間について行う。すなわち、音声区間検出を行う場合の様に、照合開始時点であるi=1を音声の始端とし、i=Iを音声の終端として照合演算をすることは意味を持たない。ワードスポッティング法では、全入力区間について音声の始終端を仮定して単語標準パターンに対する照合スコアを算出する。即ち経路判定24において行う部分類似度の累積演算は次のようになる。ここで、入力の第ｉフレーム部分と第ｊ番目の部分パターンとの部分距離を単語番号の添字kを省略してＬ(i,j)と表現し、(i,j)フレームまでの累積距離をg(i,j)と表現することにする。経路判定部２４は

【００７９】
【数１４】

$$g(i,1)=L(i,1)$$

$$g(i,j)=\min\begin{cases}g(i-2,j-1)+L(i,j)\\ g(i-1,j-1)+L(i,j)\\ g(i-1,j-2)+L(i,j-1)+L(i,j)\end{cases}$$

$$(1<i\leqq I,\ 1<j\leqq J)$$

【００８０】の演算を行い、式で示した３つの経路のうち累積距離が最小になる経路を選択する。このようにして、逐次距離を累積したのち、判定部２５では、iに対してg(i,J)が最も小さい値をとった時に、このg(i,J)を単語標準パターンの最終的な照合スコアとし、この時のiを音声の終端とする。音声の始端は、経路判定部２４の判定した経路を辿ることにより得ることができる。

【００８１】図３の経路判定部３４で行う演算は、端点固定の処理であるので、以下の様になる。

【００８２】
【数１５】

$$g(1,1)=L(1,1)$$

$$g(i,j)=\min\begin{cases}g(i-2,j-1)+L(i,j)\\ g(i-1,j-1)+L(i,j)\\ g(i-1,j-2)+L(i,j-1)+L(i,j)\end{cases}$$

$$(1<i\leqq I,\ 1<j\leqq J)$$

【００８３】（数１５）では、便宜上、音声のフレーム　50　iを番号を照合区間の始端が１、終端がＩになるように

15

つけなおしている。

【００８４】経路判定部３４は、（数１５）で示した３つの経路のうち累積距離が最小になる経路を選択する。このようにして、逐次距離を累積してゆき、j＝Jk，i＝Iとなる時点での累積距離g(i,Jk)単語ｋの照合スコアとする。判定部３５は、累積距離g(i,Jk)の最小値を与える単語ｋを求め出力する。

【００８５】以下、本発明の第２の実施例を説明する。図７は本発明の第２の実施例の構成図を示したものである。図７において、図１と同じ構成要素には同じ番号を付している。第１の実施例と異なる点は第１の照合／判定部４および単語標準パターン併合部４１であり、次にこれらで行う処理内容について詳しく説明する。　まず、単語標準パターン併合部４１で行う処理内容について説明する。部分距離計算での計算量削減のため、部分標準パターン２フレームを組とし１つにまとめる。線形判別式を用いているので、ＤＰパス上の部分距離の和を求めることは、対応するパラメータを先に加えてから部分距離を求めるのと等しい。従ってこの処理は２フレーム毎のＤＰパスを１つに固定することになる。部分標準パターン

16

【００８６】

【外１５】

$$\hat{A}_{k,n}$$

【００８７】は従来の

【００８８】

【外１６】

$$A_{k,2n-1}$$

【００８９】と

【００９０】

【外１７】

$$A_{k,2n}$$

【００９１】を１フレーム分ずらして併合して作成する。つまり従来の部分距離を（数１６）、（数１７）に示すと、

【００９２】

【数１６】

$$L_k(i,2n-1)=B_{k,2n-1}-A_{k,2n-1}\cdot X_{i-1}^{t}$$

【００９３】

【数１７】

$$L_k(i,2n)=B_{k,2n}-A_{k,2n}\cdot X_i^{t}$$

【００９４】部分距離は上記２式をまとめて、

【００９５】

【数１８】

$$\hat{L}_k(i,n)=\hat{B}_{k,n}-\hat{A}_{k,n}\cdot\hat{X}_i^{t}$$

【００９６】となり、２フレーム毎のＤＰパスを１つに固定することを条件として、従来の部分距離を求めるのと等しくなる。

【００９７】この改良により、部分距離計算に用いる特徴パラメータをＬフレーム分とすると（Ｌ＋１）／２Ｌに計算量を削減することができる。第２の実施例ではパラメータとしてＬ＝４とすると、この場合の計算量は５／８に削減できる。

【００９８】次に、第１の照合／判定部４２で行う処理内容について図面を用いて説明する。図８は第１の照合／判定部４２の処理の流れの詳細を示す構成図である。第１単語標準パターン加工部８および単語標準パターン併合部４１から得られる単語標準パターンと複数フレームバッファとの部分距離を部分距離計算部２２にて計算する。代表部分距離選択部５１で入力に対する２フレーム分の部分距離に対して、あらかじめ距離の小さい方を代表部分距離

【００９９】

【外１８】

$$\mathrm{typ}\hat{L}_k(2i,n)$$

【０１００】とすると、（数１９）となる。

【０１０１】

【数１９】

$$\mathrm{typ}\hat{L}_k(2i,n)=\min(\hat{L}_k(2i-1,n),\hat{L}_k(2i,n))$$

【０１０２】この代表部分距離について、距離累積部２３で累積し、単語全体に対する類似度を求める。その場合入力部分（Ｉフレーム）を各単語の標準時間長Jkに伸縮しながら累積する必要がある。この計算は第１の実施例と同様ＤＰ法を用いて効率よく計算できる。

【０１０３】図８では音声区間検出を行わない端点フリーの照合を行うことによりワードスポッティング法をもちいているので単語照合の処理は以下の様になる。ワードスポッティング法の場合、音声区間検出を行わないので、照合演算は音声を含む十分長い区間について行う。ワードスポッティング法では、全入力区間について音声の始終端を仮定して単語標準パターンに対する照合スコアを算出する。即ち経路判定５２において行う部分類似度の累積演算は次のようになる。ここで、入力の第ｉフレーム部分と第ｊ番目の部分パターンとの部分距離を単語番号の添字kを省略してtyp L(i,j)と表現し、(i,j)フレームまでの累積距離を

【０１０４】

17 18

【外１９】

$$\hat{g}(i,j)$$

【０１０５】と表現することにする。経路判定部５２は

$$\hat{g}(i,1) = \mathrm{typ}\,\hat{L}(i,1)$$

【０１０６】

【数２０】

$$\hat{g}(i,j) = \min \begin{cases} \hat{g}(i-1,j-1) + \mathrm{typ}\,\hat{L}(i,j) \\ \hat{g}(i-2,j-1) + \mathrm{typ}\,\hat{L}(i,j) \\ \hat{g}(i-3,j-1) + \mathrm{typ}\,\hat{L}(i,j) \end{cases}$$

$$(1 < i \leq I,\ 1 < j \leq J)$$

【０１０７】の演算を行い、式で示した３つの経路のうち累積距離が最小になる経路を選択する。このようにして逐次距離を累積した後、判定部２５ではiに対してｇ(i,J)が最も小さい値をとった時に、この ｇ(i,J)を単語標準パターンの最終的な照合スコアとし、この時のｉを音声の終端とする。音声の始端は、経路判定部５２の判定した経路を辿ることにより得ることができる。以後、第１の実施例と同様の処理を行う。

【０１０８】

【発明の効果】以上説明したように本発明は、まず間引かれたデータを用いて照合演算によってあらかじめ単語候補を絞り込んだ後に、絞り込まれた候補に対して間引きをしないデータを用いた照合を行うことにより認識性能を確保しながら計算量の削減を図るという効果があり、同一規模のハードウェアで実現する場合を考えると、従来例に比べて認識性能を保ちながら語彙数を１桁程度拡大することができる。また、環境パターンつき単語標準パターンを用いることにより音声区間の外側の区間を含めて照合をすることにより、ワードスポッティングの場合に問題となる、単語が他の単語の１部と照合することにより誤認識が生ずる部分マッチングの現象が生じないばかりか、精密な音声区間の検出も行わないため、ノイズに対して頑強な音声認識装置の実現が可能となる。

【０１０９】さらに第２の実施例においては、部分標準パターンを複数フレーム毎に併合することにより部分積の計算の削減し、かつ、ＤＰの格子点を辞書軸・入力軸それぞれ１／２に削減することで比較演算回数を大幅に削減している。

【図面の簡単な説明】

【図１】本発明の第１の実施例における音声認識装置の構成図

【図２】同実施例の構成要素である第１の照合／判定部の構成図

【図３】同実施例の構成要素である第２の照合／判定部の構成図

【図４】同実施例における音声片標準パターン作成方法の説明図

【図５】同実施例における単語標準パターンの例を示す図

【図６】同実施例における単語辞書の例を示す図

【図７】本発明の第２の実施例における音声認識装置の構成図

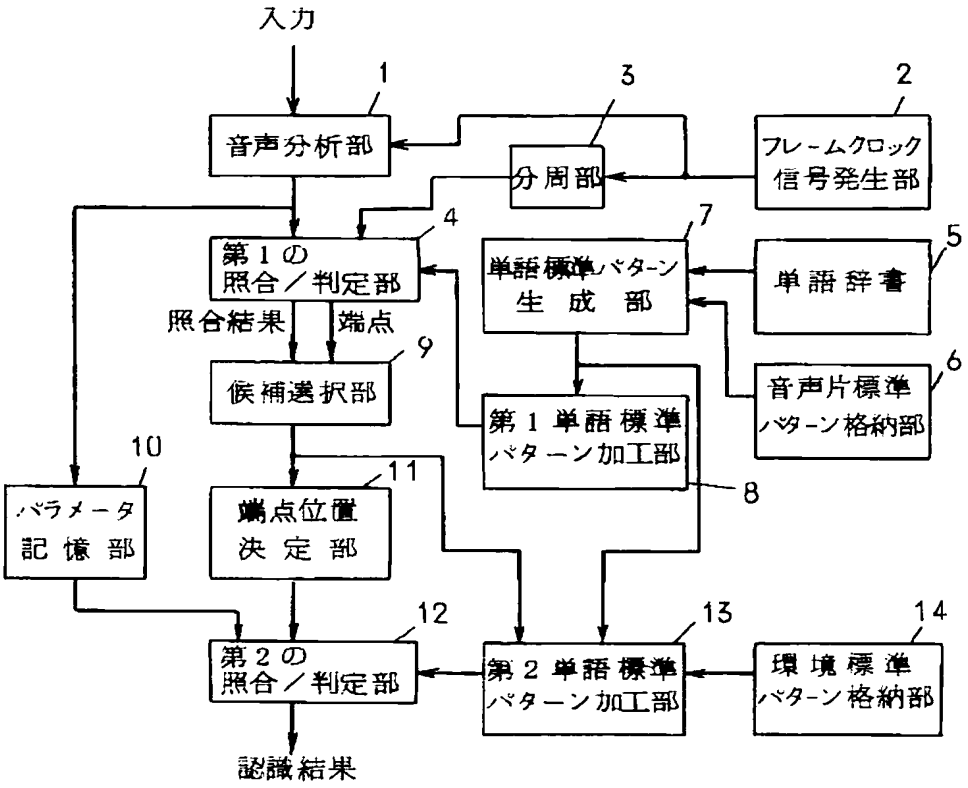
【図８】同実施例の構成要素である第１の照合／判定部の構成図

【図９】従来の音声認識装置の構成図

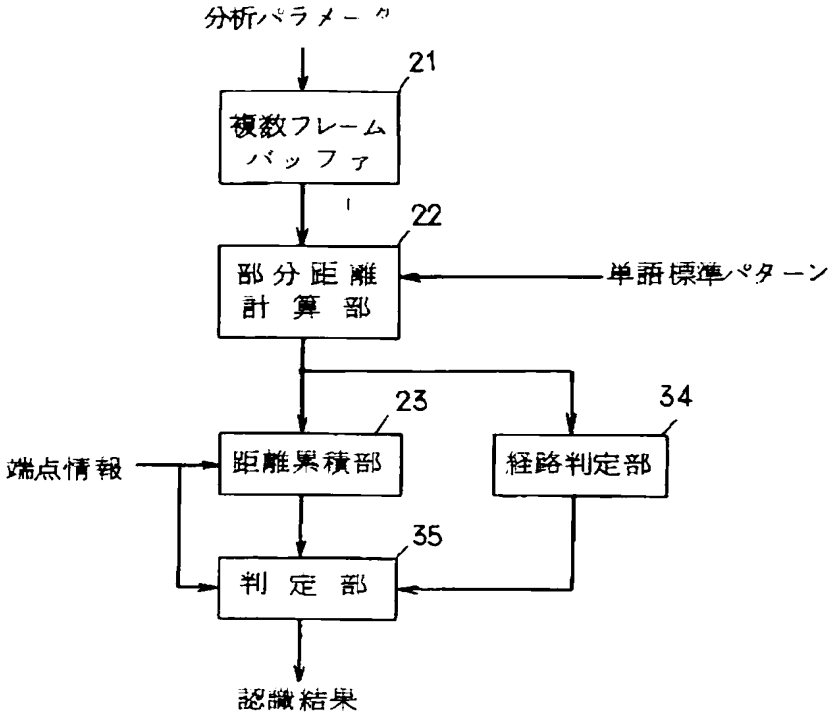
【符号の説明】

１　音響分析部
２　フレームクロック信号発生部
３　分周部
４　第１の照合／判定部
５　単語辞書
６　音声片標準パターン格納部
７　単語標準パターン生成部
８　第１単語標準パターン加工部
９　候補選択部
１０　パラメータ記憶部
１１　端点位置決定部
１２　第２の照合／判定部
１３　第２単語標準パターン加工部
１４　環境標準パターン格納部
４１　単語標準パターン併合部
５１　代表部分距離選択部

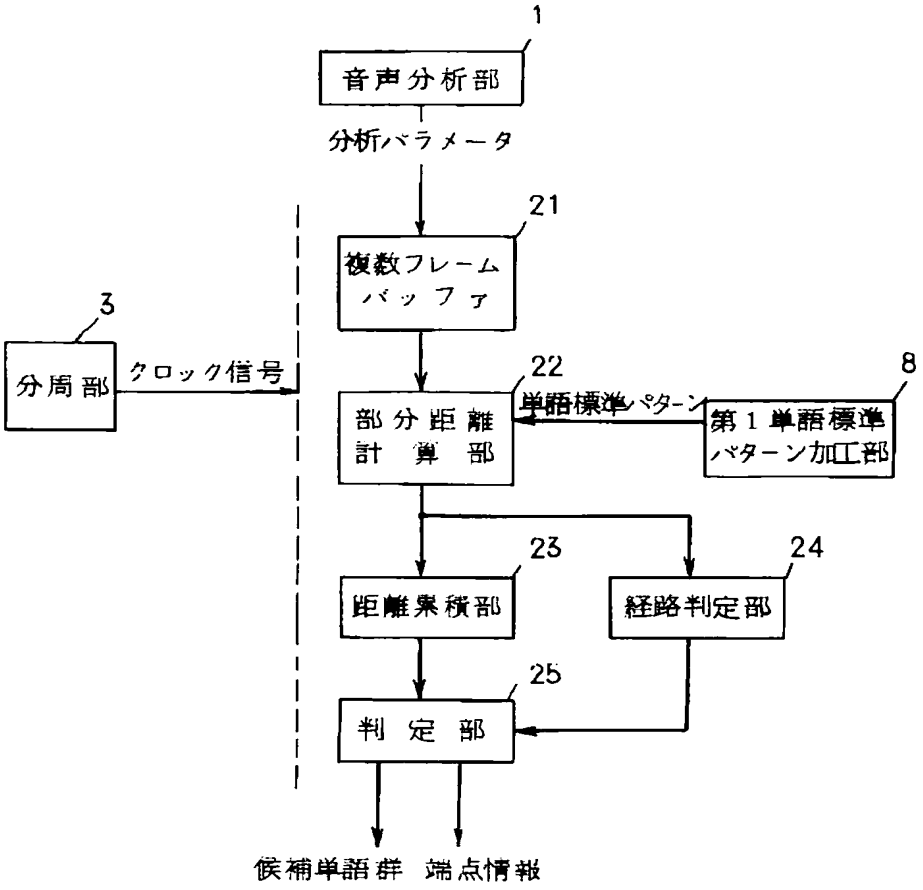
【図１】


入力
1
音声分析部
3
分周部
2
フレームクロック信号発生部
4
第１の照合／判定部
7
単語標準パターン生成部
5
単語辞書
照合結果
端点
9
候補選択部
6
音声片標準パターン格納部
第１単語標準パターン加工部
8
10
パラメータ記憶部
11
端点位置決定部
12
第２の照合／判定部
13
第２単語標準パターン加工部
14
環境標準パターン格納部
認識結果

【図３】


分析パラメータ
21
複数フレームバッファ
22
部分距離計算部
単語標準パターン
端点情報
23
距離累積部
34
経路判定部
35
判定部
認識結果

【図２】

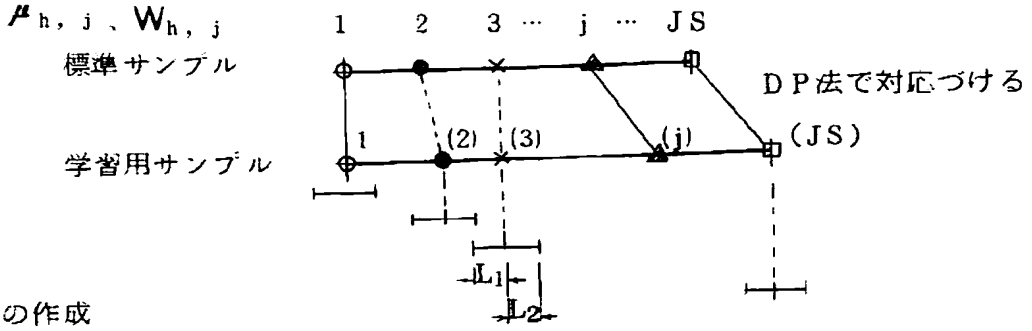



【図５】

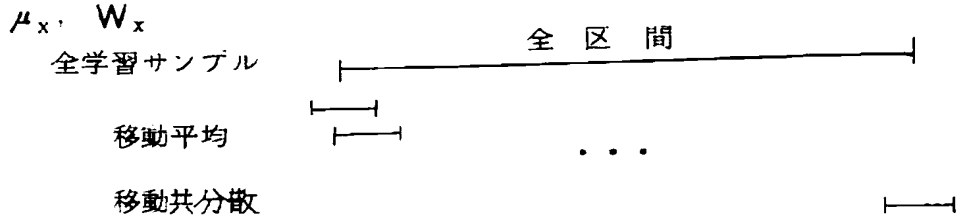
| 単語 | 音声片表記 |
|---|---|
| あさひ | ／a／／s a／／h i／ |
| いろがみ | ／i／／r o／／g a／／m i／ |
| おおさま | ／o／／o／／s a／／m a／ |
| ・ | ・ |
| ・ | ・ |
| ・ | ・ |

【図４】

音声片標準パターンの作成

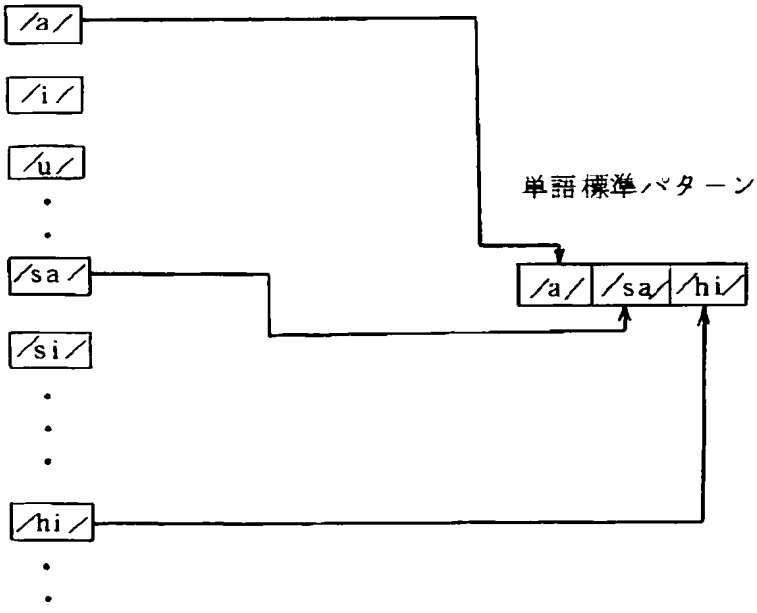



周囲パターンの作成

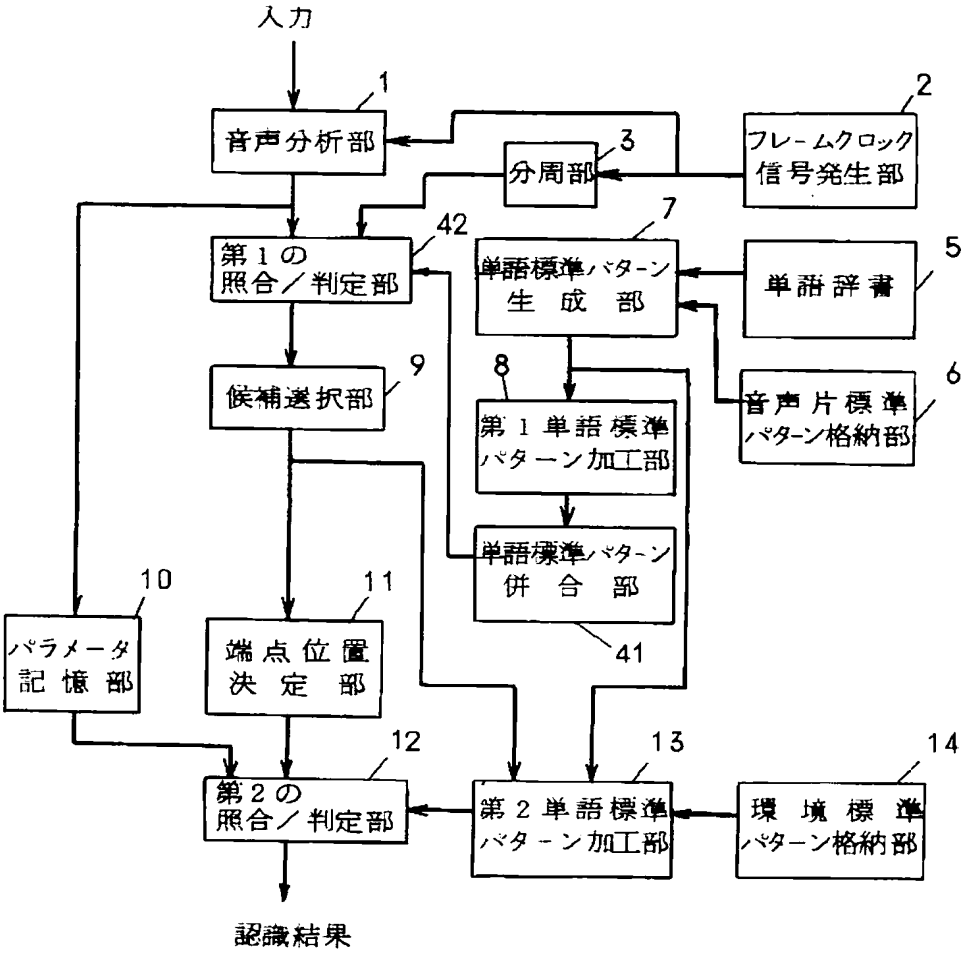



【図６】

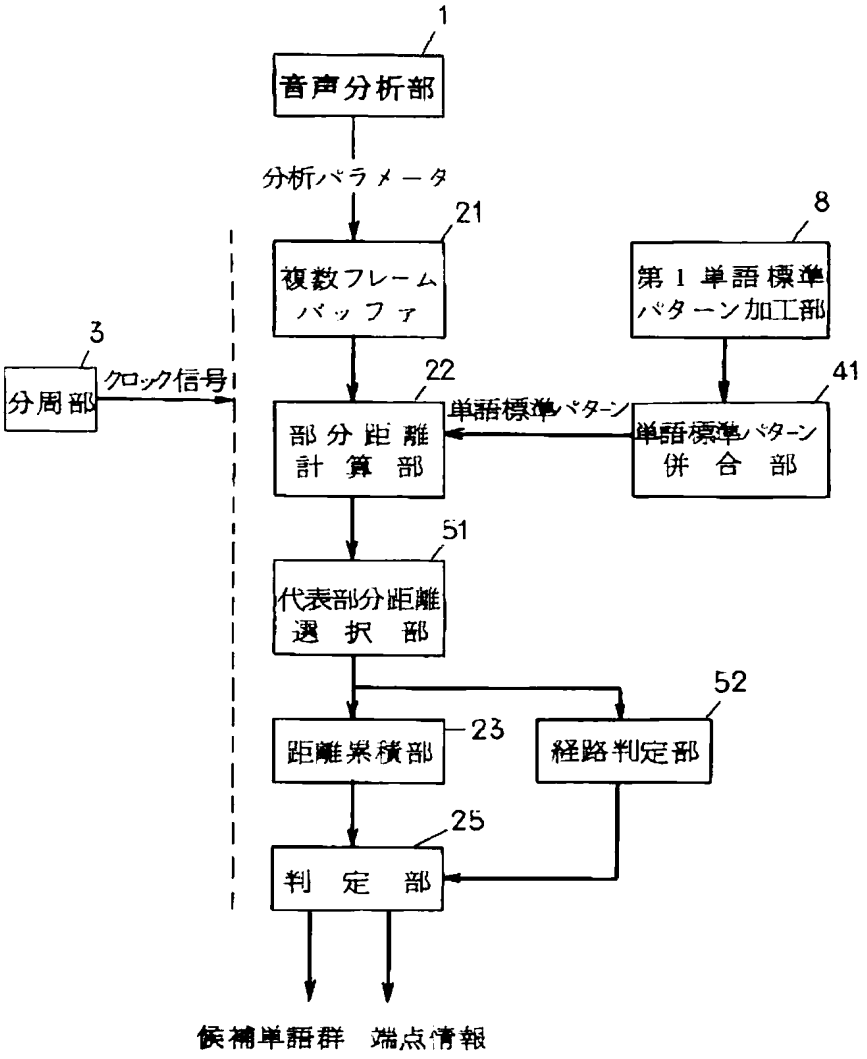
音声片標準パターン



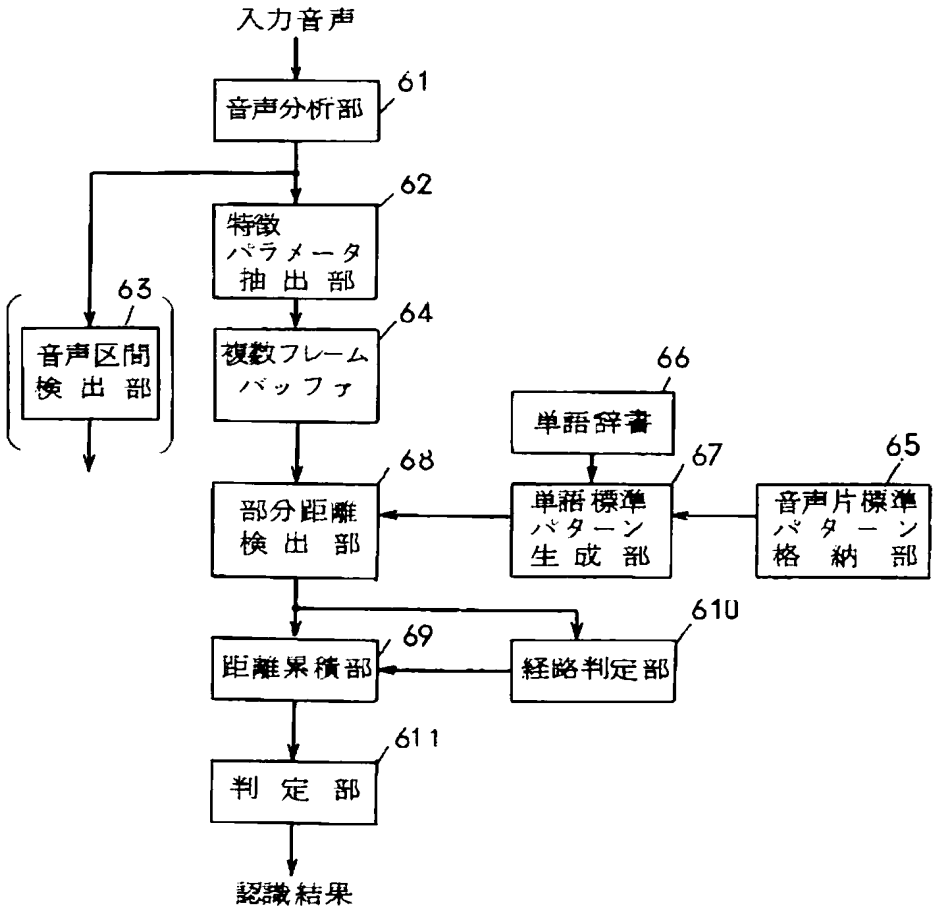
【図７】


入力
1
音声分析部
2
フレームクロック信号発生部
3
分周部
42
第１の照合／判定部
7
単語標準パターン生成部
5
単語辞書
9
候補選択部
8
第１単語標準パターン加工部
6
音声片標準パターン格納部
単語標準パターン併合部
41
10
パラメータ記憶部
11
端点位置決定部
12
第２の照合／判定部
13
第２単語標準パターン加工部
14
環境標準パターン格納部
認識結果

【図８】


1
音声分析部
分析パラメータ
21
複数フレームバッファ
8
第１単語標準パターン加工部
3
分周部
クロック信号
22
部分距離計算部
単語標準パターン
単語標準パターン併合部
41
51
代表部分距離選択部
52
距離累積部
23
経路判定部
25
判定部
候補単語群
端点情報

【図９】




フロントページの続き

| (51) Int.Cl.[5] | 識別記号 | 庁内整理番号 | F I | 技術表示箇所 |
|---|---|---|---|---|
| G10L 3/00 | | J 9379－5H | | |

(72)発明者　平岡　省二
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内